# DeepInception: Hypnotize Large Language Model to Be Jailbreaker

*Year: 2023 | Citations: 271 | Authors: Xuan Li, Zhanke Zhou, Jianing Zhu*

---

## Abstract

Large language models (LLMs) have succeeded significantly in various applications but remain susceptible to adversarial jailbreaks that void their safety guardrails. Previous attempts to exploit these vulnerabilities often rely on high-cost computational extrapolations, which may not be practical or efficient. In this paper, inspired by the authority influence demonstrated in the Milgram experiment, we present a lightweight method to take advantage of the LLMs' personification capabilities to construct $\textit{a virtual, nested scene}$, allowing it to realize an adaptive way to escape the usage control in a normal scenario. Empirically, the contents induced by our approach can achieve leading harmfulness rates with previous counterparts and realize a continuous jailbreak in subsequent interactions, which reveals the critical weakness of self-losing on both open-source and closed-source LLMs, $\textit{e.g.}$, Llama-2, Llama-3, GPT-3.5, GPT-4, and GPT-4o. The code and data are available at: https://github.com/tmlr-group/DeepInception.